DOCUMENT RESUME

ED 090 308 TM 003 596

AUTHOR	McMullen, David W.
TITLE	Realism Training Through Decision-Theoretic Testing.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C.
PUB DATE	74
GRANT	OEG-2-72-2B071
NOTE	20p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, Illinois, April, 1974)
EDRS PRICE	MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS	Complexity Level; *Confidence Testing; Decision Making; *Feedback; Junior High School Students; Peer Relationship; Probability Theory; *Realism; Testing; *Training
IDENTIFIERS	*Decision Theoretic Testing

ABSTRACT

Decision-theoretic testing is used to explore whether students can improve their realism, i.e. congruence between reported and true probabilities. Randomized sets of math problems were presented at computer terminals to 49 seventh graders from two classes (high/low achievers) over a period of three weeks. The subject assigned values to each of four alternatives for each problem. Two groups worked individually, one with realism feedback. A third group, also with feedback, worked in teams of size three. Only the individual-study feedback groups in the low class improved on Posttest I. On Posttest II, with harder items and peer interaction, nearly all subjects overvalued information. Deviation from realism consistently increased with task difficulty and the direction of deviation was nearly always toward overconfidence. (Author)

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

ED 090308
TM 003 596

REALISM TRAINING THROUGH DECISION-THEORETIC TESTING

David W. McMullen
SUNY, Stony Brook

Paper presented at the 1974
Annual Meeting of the
American Educational Research Association
April, 1974

Session 13.22

ABSTRACT

Decision-theoretic testing is used to explore whether students can improve their realism, i.e. congruence between reported and true probabilities. Randomized sets of math problems were presented at computer terminals to 49 seventh graders from two classes (high/low achievers) over a period of three weeks. S assigned values to each of four alternatives for each problem. Two groups worked individually, one with realism feedback. A third group, also with feedback, worked in teams of size three. Only the individual-study feedback groups in the low class improved on Posttest I. On Posttest II, with harder items and peer interaction, nearly all Ss overvalued information. Deviation from realism consistently increased with task difficulty, and the direction of deviation was nearly always toward overconfidence.

REALISM TRAINING THROUGH DECISION-THEORETIC TESTING [1]

Applications of decision theory to testing (de Finetti, 1965; Albert and Massengill, 1966; Wang and Stanley, 1970; Echternacht, 1972) have provided methods by which to assess knowledge without fostering confusion between belief and certainty (Savage, 1971). For educators who assert that it is important for students to know what they don't know, decision-theoretic testing offers an opoortunity to measure a dimension of student achievement called realism, the congruence of reported probabilities with true probabilities. The degree of student realism reflects the degree to which the student recognizes the worth of what he knows. Lack of realism indicates that the student either overvalues or undervalues his information and reasoning (Shuford, 1973).

This experiment examined the degree of realism in student assessment tendencies and considered whether realism can be improved by systematic training based on decision-theoretic testing.

METHOD

Materials

A set of panels, called a "SCoRule" (Shuford, 1970), was used to record probability assessments in the form of logarithmic values. Each panel contained a black line (representing probability) below a curved logarithmic scale ranging from 0 to 100. With one panel for each alternative of a multiple choice question, the SCoRule allows varying lengths of the black line on each panel to be exposed while the combined length of all exposed black lines is kept constant (i.e. equal to 1.00). The point on the curved scale that corresponds to the end of an exposed line provides the logarithmic value of the probability represented by that line segment. The logarithmic transformation has been shown to maximize the score on an item for which reported probability agrees with true probability (Shuford et al, 1966).

All questions had four choices and came from the mathematical problem-solving sections of the Iowa Test of Basic Skills. Forms 1-4 were pooled and divided into 15 levels of difficulty on the basis of item analysis information provided with the test. Nine problem sets were formed by assigning the first item at each level to the problem set for the first training session, the second item at each level to the problem set for the second session, etc. Form 5 was administered as the pretest and Form 6 as the first posttest. A second posttest consisted of the last 21 items in Form 5, which no S reached

during the pretest.

A SCoRule Procedures Test (McMullen, 1973) was administered at the time of the pretest after pre-training with sample, non-experimental problems.

Subjects

Two seventh grade classes from a suburban junior high school were bussed to the Stony Brook Computer-Assisted Instruction Laboratory during their regular 45-minute math periods. One class, taught by the department chairman, was considered higher in achievement level than the other class. A total of 54 Ss participated, but because of absences and technical problems only 49 provided adequate data for analysis, 26 in the high-achieving class and 23 in the low-achieving class.

Experimental task and realism graph

Nine training sessions, each approximately 25 minutes in length, occurred during a three-week period. Using scope-and-keyboard terminals linked to an IBM 1500 Instructional System, Ss registered SCoRule (log) values opposite each of four alternatives for as many items as time permitted. The computer converted log values to probabilities and required S to retype SCoRule values if the probability equivalents did not sum to 1.00 within a tolerance of .04. Otherwise the computer underlined the SCoRule value given by S to the correct alternative, added that value to the cumulative score, and displayed the score.

At the end of each session, some Ss viewed a realism graph.

On the horizontal axis were levels of reported probability between 0.00 and 1.00. On the vertical axis were proportions, also from 0.00 to 1.00, representing the proportion of times a given probability level was associated with a correct alternative out of the total number of times the probability level was used (Fig. 1).

- -

Insert Figure 1 about here

- -

For example, ten four-alternative items result in 40 reported probabilities. If 5 of the 40 are .60 and 3 of the 5 are associated with a correct alternative, the value on the vertical axis corresponding to .60 on the horizontal axis would also be .60. When values on the vertical axis match probability levels on the horizontal axis, the result is a 45-degree line (slope of 1.00), indicating perfect realism. A slope less than 1.00 indicates a tendency to overvalue one's knowledge (e.g. .60 associated with a correct alternative only 2 out of 5 times), and a slope greater than 1.00 a tendency to undervalue (e.g. .60 correct 4 out of 5 times). (For a technical discussion, see Brennan, 1973.)

The computer-generated graph differed in three respects from Fig. 1: (1) the vertical axis was altered so that the observed realism line appeared above rather below the perfect realism line when S was overconfident, above when S was underconfident, (2) axes were unlabeled, and (3) verbal feedback appeared with the graph. If the observed realism line fell within 5 degrees of perfect realism, "realistic" appeared;

otherwise, "overconfident" or "underconfident." Allowing a tolerance of 5 degrees for defined realism was the result of a simulation procedure and subjective judgment (McMullen, 1973).

Treatments and procedure

Three groups were formed in each class. Each teacher assigned one third of his class to a team-study group keeping each team congenial and representative of the class (high achievers balanced by low achievers). The remainder were divided on the basis of number of pretest items completed and score on the SCoRule Procedures Test into four blocks (high-achiever/high SCoRule-familiarity to low-achiever/low familiarity) and randomly assigned by blocks to the individual study treatments.

The individual-study/no graph (IN) treatment received computer support in the form of immediate feedback but was not told explicitly whether realism was attained. The individual-study/graph (IG) treatment received realism graphs at the end of each session. The team-study/graph (TG) treatment was identical to Treatment IG except that S worked in three-member teams, one of which in each class was male.

Following one classroom pre-training session and one pre-training session in the CAI lab, the paper-and-pencil pretest was administered in the classroom, followed by the SCoRule Procedures Test. Posttests, also administered in the classroom, occurred the week following the ninth training session. Both the pretest and Posttest I were taken individually, while Posttest II was administered to Ss working as teams of three.

Teachers formed the teams, keeping Treatment TG teams the same. Though encouraged to consult with team members, S recorded SCoRule values separately on Posttest II.

During training S was advised to begin a problem with the SCoRule indicating "no knowledge," i.e. all values the same. Values could then be raised or lowered depending on S's knowledge. When a graph appeared, E interpreted the results with statements such as "'Overconfident' ('Underconfident') means you are placing too many (few) points on what you consider the best alternative."

Results

Angle of Deviation from Perfect Realism

The extent to which observed realism deviated from perfect realism (see Fig. 1) decreased from pretest to Posttest I by 2 degrees and variability decreased on both posttests, but an increase in the angle of deviation (realism loss) occurred on Posttest II. On Posttest I Tables 1 and 2 show that realism

- -

Insert Tables 1 and 2 about here

- -

gain occurred in Treatments IN and IG and in the low-achieving class, but Wilcoxen signed-rank tests failed to disclose significance ($z = 2.64$, 2.83, and 2.64).

During training sessions, mean deviation angles for the total sample remained between the means of Posttest I and Posttest II, with marked differences between the high class and low class. The mean angle for the high class hovered consistently

near 10 degrees, but for the low class fluctuated between 10 and 30 degrees.

Deviation angles throughout the experiment were closely related to item difficulty levels. Item difficulty was computed by considering the most highly valued alternative as the response S could have given on a conventional test. If several alternatives had the same high value, one was randomly chosen in order to estimate S's response. Tests and training sessions for which mean deviation angles were relatively high tended to contain items on which Ss performed relatively poorly.

Additional evidence of a relationship between realism and item difficulty is shown in Figure 2. Because of random

- -

Insert Figure 2 about here

- -

presentation of training session items, each problem set could be divided into two comparable subsets: items from the easiest and the hardest seven levels (see above, Materials). Except for the third session, deviation from realism was consistently greater for hard than it was for easy items.

Defined Realism and Overconfidence

Feedback in the form of a realism graph was based on defined realism, i.e. a deviation angle no greater than five degrees. Table 3 shows the number in each class and treatment group who

- -

Insert Table 3 about here

- -

attained defined realism on the pretest and Posttest I, and

those who deviated in either direction. For the total sample the proportion who attained defined realism rose from .25 to .33, but the increase came entirely from the low class. The only treatment group in the low class with a higher posttest proportion studied individually and received the graph.

On Posttest II the entire sample showed overconfidence except for two in the low class, both underconfident, one in each individual study treatment. Throughout the experiment, the preponderance of those who deviated from realism exhibited overconfidence, particularly in the high class. During training sessions underconfidence occurred only 3 times in the high class, 13 times in the low class, 4 times in Treatment IN, 12 times in Treatment IG, and not once in Treatment TG.

DISCUSSION

Student Characteristics

Assessment tendencies varied widely in stability but almost uniformly demonstrated overconfidence in the worth of information. The fluctuations of the low class contrasted sharply with the steady performance of the high class with respect to deviation from realism. Both classes, however, characteristically deviated in the same direction. Despite opportunities to gain familiarity with the process of assessing subjective probabilities, students overvalued information toward the end of the experiment in about the same way they did at the beginning.

A stable relationship appeared between realism and problem difficulty, suggesting that distortion tendencies are sensitive

to changes in achievement level and the degree of risk represented in a task. Furthermore, the relationship does not appear to be linear because fluctuations in the deviation between observed and perfect realism were much larger for the low class (high task difficulty) than for the high class (low task difficulty).

Though high task difficulty was associated with disproportionate information distortion, the direction of that distortion was not uniformly toward overconfidence. In fact, overconfidence was more characteristic of the class that had relatively less difficulty with the problems. Nearly all of the few students who were underconfident during the experiment performed at relatively low levels.

A factor that appeared to contribute to overconfidence, besides high achievement level, was peer interaction. Both the teamed treatment group and the teamed posttest indicated that group dynamics promote a tendency to ignore the limitations of what one knows. A group, for example, may introduce distortion by subtly undermining respect for one's own judgment or by bringing extraneous personality factors into the decision-making process. Though it is employed in this study primarily as an information-theoretic construct, not to be confused with an exaggerated sense of self-worth, overconfidence appears to be closely linked to interpersonal and affective characteristics that dispose an individual to attach undue weight to information.

Realism Training

The data supports a tentative conclusion that, despite the lack of overall dramatic gains, realism training has beneficial effects. In particular, low achievers who viewed realism graphs increased sharply in the proportion who attained defined realism, while low achievers without the graph did not. It should also be noted that a general reduction in the variability of information distortion also occurred for the sample as a whole following training.

The study suggests that realism gain may require a combination of two conditions: feedback and high problem difficulty. Students may be more likely to discover long-established tendencies to distort the value of information when uncertainty is at a high level rather than only moderate. Moreover, feedback is apparently needed in order to form an intuitive awareness of what it means to be realistic.

Time limitations may be responsible for the lack of substantial overall gain. The modification of tendencies fostered by our culture probably requires a period of months, not weeks. In fact, it is surprising that nine 20-25 minute sessions, with an average of 7-8 items per session, provided enough training to be associated with any change in observed realism. In addition, some students were occasionally not able to view their graphs because of the rush to board the bus.

Contrary to expectations, team problem-solving did not appear to support realism training. Especially with relatively difficult

problems (Posttest II), students found realism hard to achieve while interacting with their peers. Even without team study, however, students clearly tended to exaggerate the value of their information. To determine whether realism training can alter those tendencies is a subject for continued investigation.

FOOTNOTE

[1]The research reported in this paper was supported by Grant No. OEG-2-72-2B071, Office of Education, U.S. Dept. of HEW (McMullen, 1973).

REFERENCES

Brennan, R. Manual for DEC-TEST: A FORTRAN IV computer program for decision-theoretic test scoring and the analysis of item data. Appendix to Final Report Grant OEG-2-2-2B118, Office of Education (mimeo). Stony Brook, N.Y.: SUNY, Stony Brook, 1973.

deFinetti, B. Methods for discriminating levels of partial knowledge concerning a test item. British Journal of Mathematical and Statistical Psychology, 1965, 18, 87-123.

Echternacht, G. The use of confidence testing in objective tests. Review of Educational Research, 1972, 42, 217-236.

McMullen, D. Realistic self-assessment of knowledge and competence. Final Report, Grant OEG-2-72-2B071, Office of Education (mimeo). Stony Brook, N.Y.: SUNY, Stony Brook, 1973.

Savage, L. Elicitation of personal probabilities and expectations. Journal of the American Statistical Association, 1971, 66, 783-801.

Shuford, E., Albert, A., & Massengill, H. Admissible probability measurement procedures. Psychmetrika, 1966, 31, 125-145.

Wang, M. & Stanley, J. Differential weighting: A review of methods and empirical studies. Review of Educational Research, 1970, 40, 663-705.

FIGURE CAPTIONS

Fig. 1. Realism graph, showing relationship between reported (observed) probabilities and proportion of associated successes. Numbers beside points give proportion bases, i.e. the number of time P_i was used. For example, .20 was used three times and correct once. (After Brennan, 1973, p. A-8)

Fig. 2. Angle (in degrees) between perfect and observed realism during training sessions for easy items (solid line) and hard items (broken line).

Table 1

Mean Angle of Deviation from Perfect Realism on Pretest and Posttests By Treatment Groups in Degrees (Standard Deviation in Parentheses)

Treatment	Pretest	First Posttest	Realism Gain/Loss	Second Posttest	Realism Gain/Loss over Pretest
IN (N=15)	14.9 (11.0)	11.1 (7.2)	+3.8 (+3.8)	17.4 (6.2)	-2.5 (+4.8)
IG (N=18)	14.4 (9.9)	10.8 (9.1)	+3.6 (+ .8)	18.1 (7.7)	-3.7 (+2.2)
FG (N=16)	10.6 (12.6)	12.2 (9.4)	-1.6 (+3.2)	22.2 (9.6)	-11.6** (+3.0)
Total Sample (N=49)	13.3 (11.3)	11.3 (8.7)	+2.0 (+2.6)	19.3 (8.3)	-6.0** (+3.0)

Table 2

Mean Angle of Deviation from Perfect Realism on Pretest and Posttests By Classes in Degrees (Standard Deviation in Parentheses)

Class	Pretest	First Posttest	Realism Gain/Loss	Second Posttest	Realism Gain/Loss over Pretest
High (N=26)	9.2 (7.5)	9.1 (5.9)	+ .1 (+1.6)	16.2 (4.4)	-7.0** (+3.1)
Low (N=23)	18.0 (13.0)	13.8 (10.5)	+4.2 (+2.5)	23.5 (10.3)	-5.5 (+2.7)

** $p < .01$, Wilcoxen signed-ranks test

Table 3

Distribution in Three Categories of Observed Realism
By Treatment Groups within Classes
on Pretest and First Posttest

	Treatment Group	Pretest			Posttest I		
		Under-conf.	Real-istic*	Over-conf.	Under conf.	Real-istic*	Over conf.
High Class	IN	1	1	5	0	2	5
	IG	0	2	7	0	2	7
	TG	0	6	4	0	5	5
	All	1	9	16	0	9	17
Low Class	IN	1	1	6	2	1	5
	IG	2	0	7	1	5	3
	TG	0	2	4	0	1	5
	All	3	3	17	3	7	13
Total Sample		4	12	33	3	16	30

*observed realism within five degrees of perfect realism

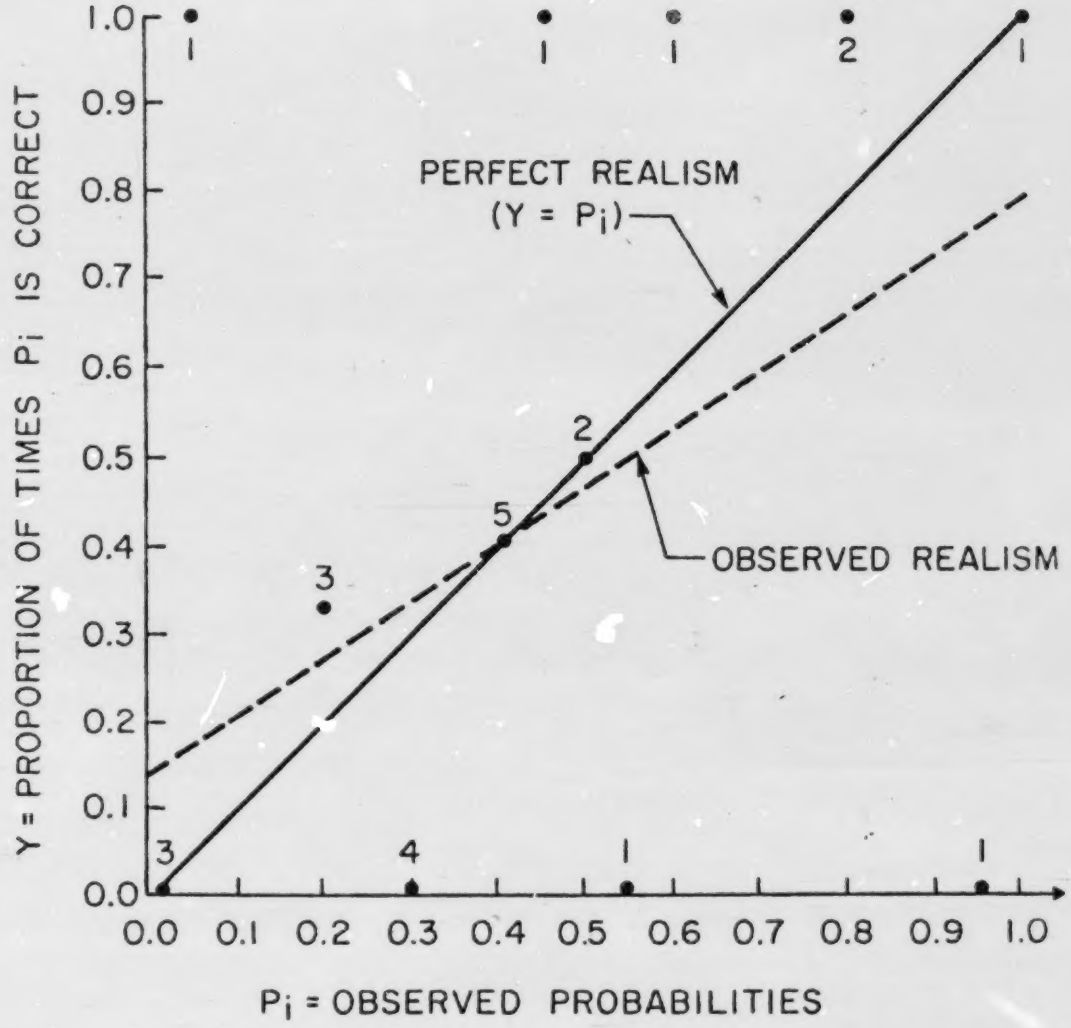
Y = PROPORTION OF TIMES Pi IS CORRECT
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
PERFECT REALISM
(Y = Pi)
OBSERVED REALISM
1
1
1
2
1
2
5
3
3
4
1
1
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Pi = OBSERVED PROBABILITIES

Figure 1.

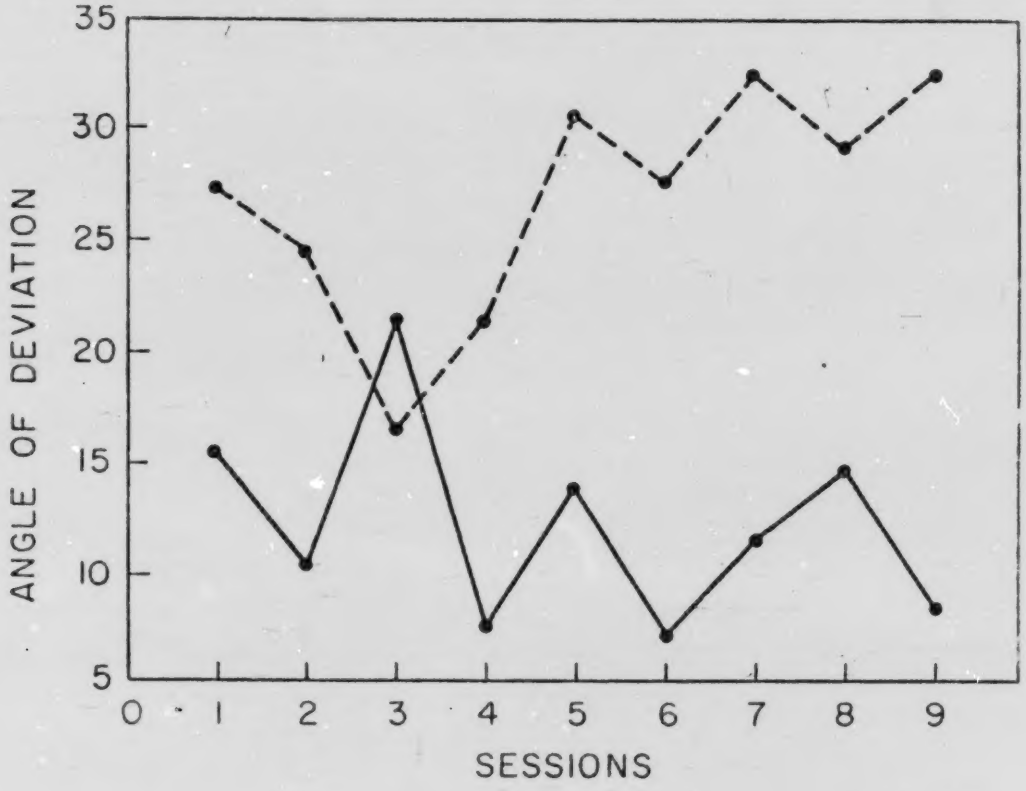

Figure 2